Ordering Monte Carlo Markov Chains

By
Antonietta Mira and Charles J. Geyer

Technical Report No. 632

School of Statistics
University of Minnesota

March 31, 1999

Abstract

Markov chains having the same stationary distribution π can be partially ordered by performance in the central limit theorem. We say that one chain is at least as good as another in the *efficiency* partial ordering if the variance in the central limit theorem is at least as small for every $L^2(\pi)$ functional of the chain. Peskun partial ordering implies efficiency partial ordering [25, 30].

Here we show that Peskun partial ordering implies, for finite state spaces, ordering of all the eigenvalues of the transition matrices, and, for general state spaces, ordering of the suprema of the spectra of the transition operators. We also define a *covariance* partial ordering based on lag one autocovariances and show that it is equivalent to the efficiency partial ordering when restricted to reversible Markov chains. Similar but weaker results are provided for non-reversible Markov chains.

Keywords: Peskun ordering, Eigenvalues, Spectral decomposition, Non-reversible kernels.

1 Introduction

In the past decade, Markov chain Monte Carlo (MCMC) has become widely used in statistics for calculating by computer simulation quantities with no analytic expression. The basic idea for calculating the expectation of a function f of a random variable X having distribution π

$$E_\pi[f(X)] = \int f(x)\pi(dx) = \mu. \tag{1}$$

is to run a Harris recurrent Markov chain $X_1, X_2, \ldots$ having π as its stationary distribution. The sample average

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

is then used as the Monte Carlo approximation of μ. The law of large numbers (LLN) for Markov chains [22, Proposition 17.1.6 and Theorem 17.1.7] implies that $\hat{\mu}_n$ converges almost surely to μ, which justifies MCMC just like the ordinary LLN justifies ordinary Monte Carlo.

For complicated probability models, we typically use MCMC because it is easy to find Markov chains having the required stationary distribution and impossible to find an ordinary independent-sample Monte Carlo scheme for that distribution. The class $\mathcal{P}$ of Markov chains having π as their stationary distribution can be quite large. Thus the question naturally arises, which is the best Markov chain in $\mathcal{P}$ for MCMC purposes? To decide that we need a performance criterion. What do we mean by "best"? Here, as elsewhere in statistics, the most useful criterion is variance in the central limit theorem (CLT). Under certain regularity conditions the CLT for Markov chains

$$\sqrt{n}\,(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2) \tag{2}$$

holds and gives an approximation of the Monte Carlo error just like that given by the ordinary CLT for ordinary Monte Carlo.

The variance σ^2 in the CLT depends on the function f being integrated and it also depends on the particular Markov chain we are using. Markov chains differ in having different *transition kernels*

$$P(x, A) = \Pr(X_n \in A | X_{n-1} = x).$$

We note this dependence by writing the variance σ^2 in the CLT as $v(f, P)$. If for a particular function f and transition kernel P the CLT (2) does not

hold, then we define $v(f, P)$ to be ∞. In this paper we will use the symbol P to refer to both the Markov chain updated using P and the corresponding transition kernel.

The set of reversible transition kernels with respect to π is a subset of $\mathcal{P}$. In the first part of this paper we restrict our attention to this subset while in the second part (Section 5 onwards) we try to extend the results to non-reversible transition kernels. The difficulty lies in the fact that, for non-reversible transition kernels, we cannot use spectral theory or classical functional analysis tools. Moreover, intuition often fails to support the reasoning or, even worse, intuition can be misleading.

Contrary to what is often done in classical statistical inference when looking for minimum variance estimates, we do not assume any prior knowledge of the function whose expectation we want to evaluate. Thus, given two Markov chains P and Q in $\mathcal{P}$, we say that P is more efficient than Q if $v(f, P) \leq v(f, Q)$ for all functions f that obey the CLT (efficiency ordering). In Section 3 we study a partial ordering that implies the efficiency ordering (Peskun ordering).

In Section 4 we provide a necessary and sufficient conditions for a Markov chain to be more efficient than another (covariance ordering).

Section 3.1 shows that, in finite state spaces, the Peskun ordering induces an ordering on the eigenvalues of the corresponding transition matrices but not on the absolute values of the eigenvalues. The distinction is quite relevant: fast convergence to stationarity in total variation distance is reached by having small eigenvalues in absolute value, while small asymptotic variance of MCMC estimates is achieved by having small eigenvalues. Therefore, unless the operators used are positive (i.e. have positive eigenvalues or a positive spectrum), we are faced with conflicting goals.

In Section 3.2 we try to extend to general state spaces the result on ordering the eigenvalues. Here the difficulty lies in the fact that we cannot talk about eigenvalues anymore but we need to introduce the concept of a spectrum.

In Section 7 some non-reversible Markov chains are analyzed using the tools developed in Section 5.

The last two sections of the paper are dedicated to the comparison of the performance of reversible and non-reversible Markov chains. In Section 8 we try to answer the following question: "given a non-reversible Markov chain when can we find reversible one with the same stationary distribution which is at least as efficient?" We do not have a definite answer to this question but give guidelines and intuitions on how to address the problem. In Section 9 we make some considerations regarding the comparison of performances of

Markov chains taking into account the amount of "labor" expended to run them. From a practitioner's point of view this, rather than efficiency comparison on a sweep-by-sweep basis, is what matters.

2 Preliminaries

In this section we set up the notation needed in the sequel and review some of the theory on MCMC. Let $\{X_n\}_{n=1}^{\infty}$ denote a Markov chain P in the class $\mathcal{P}$, that is, such that

$$\pi(A) = \int P(x, A)\pi(dx) \tag{3}$$

for all measurable sets A. Assume that the initial distribution of the chain is equal to the stationary distribution π. This gives us a *stationary Markov chain*, that is, the distribution of X_n does not depend on n. Given a nonzero measure on the state space, φ, a Markov chain is said to be *φ-irreducible* if, for any point x and any measurable set A such that $\varphi(A) > 0$, there exists an integer n such that $P^n(x, A) > 0$. For a φ-irreducible Markov chain, conditional on the starting position $X_1 = x$, $\hat{\mu}_n$ converges almost surely to μ for π-almost all x (Birkhoff ergodic theorem, [10]). If furthermore the chain is *Harris recurrent*, then almost sure convergence holds for any initial distribution (Proposition 17.1.6 [22]). The same principle is true for the CLT. If the chain is Harris recurrent the CLT holds for all initial distributions if it holds for the stationary distribution. So, without loss of generality, we can work with stationary Markov chains when dealing with the strong law of large number or the CLT.

For general state spaces several conditions have been stated that guaranty the existence of a CLT such as uniform, geometric ergodicity or other mixing conditions [29, 22, 27]. In Section 3.2 a general sufficient condition for CLT will be given. A detailed discussion on this issue is available in [29] and [5]. The variance in the CLT, $v(f, P)$, is the limit, as n tends to infinity, of

$$\begin{aligned}
\sigma_n^2 &= n \operatorname{Var}_\pi[\hat{\mu}_n] \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}_\pi[f(X_i), f(X_j)] \\
&= \frac{1}{n} \sum_{i=1}^{n} \operatorname{Var}_\pi[f(X_i)] + \frac{2}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{Cov}_\pi[f(X_i), f(X_j)].
\end{aligned}$$

Since all the expectations are taken under the stationary distribution π, $\mathrm{Var}_\pi[f(X_n)]$ does not depend on n and $\mathrm{Cov}_\pi[f(X_n), f(X_{n+k})]$ does not depend on n for fixed k. Hence

$$\sigma_n^2 = \mathrm{Var}_\pi[f(X_i)] + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\,\mathrm{Cov}_\pi[f(X_i), f(X_{i+k})]. \tag{4}$$

To simplify the notation let

$$\gamma_0 = \mathrm{Var}_\pi[f(X_i)]$$

and

$$\gamma_k = \mathrm{Cov}_\pi[f(X_i), f(X_{i+k})] \tag{5}$$

which is the lag k autocovariance of the stationary time series $\{f(X_n)\}_{n=1}^{\infty}$. If the CLT holds, we might expect the limiting variance to be the limit of (4) as $n \to \infty$. If this is the case we have ([11], Chapter 3)

$$v(f, P) = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k. \tag{6}$$

The asymptotic variance in the CLT given in (6) defines our criterion for ranking transition kernels.

A transition kernel is *reversible* with respect to π if, for all bounded functions f and g,

$$\iint f(y)g(x)\pi(dx)P(x, dy) = \iint f(x)g(y)\pi(dx)P(x, dy). \tag{7}$$

Condition (7) is also known as the detailed balance condition. Let $\mathcal{D}$ be the class of transition kernels for which (7) is satisfied. If (7) holds, then by taking $g(x) = 1$ we have

$$\iint f(y)\pi(dx)P(x, dy) = \iint f(x)\pi(dx)P(x, dy) = \int f(x)\pi(dx)$$

and thus π is the stationary distribution for P so that $\mathcal{D} \subset \mathcal{P}$.

A transition kernel $P \in \mathcal{P}$ defines an *operator* on the Hilbert space $L^2(\pi)$ of square integrable functions with respect to π. The operator corresponding to P is described by the way P acts on a generic element $g \in L^2(\pi)$:

$$(Pg)(x) = E[g(X_t)|X_{t-1} = x] = \int g(y)P(x, dy). \tag{8}$$

The inner product on $L^2(\pi)$ is

$$(f,g) = \int f(x)g(x)\pi(dx) = E_\pi[f(x)g(x)].$$

Let $L_0^2(\pi) = \{ g \in L^2(\pi) : \int g d\pi = 0 \}$ be the subspace of L^2 of zero mean functions. One of the reasons we often restrict to this subspace is that, for zero mean functions f and g, the inner product (f,g) is equal to the covariance of f and g under stationarity

$$(f,g) = \mathrm{Cov}_\pi[f(x), g(x)].$$

The other reason why $L_0^2(\pi)$ is relevant for our purposes is related to its spectrum and will become clear in Section 3.2. Another way to describe $L_0^2(\pi)$ is the subspace of $L^2(\pi)$ orthogonal (inner product equal to zero) to the constant functions.

Let P^* be the *adjoint* of P, that is, the unique operator such that

$$(f, Pg) = (P^* f, g), \qquad \forall f, g \in L^2(\pi).$$

An operator is said to be *self-adjoint* if $(f, Pg) = (Pf, g)$, for any function f and g in $L^2(\pi)$. A transition kernel is reversible if and only if the corresponding operator is self-adjoint.

An operator is positive on $L^2(\pi)$, $P \geq 0$, if

$$(Pf, f) = \iint f(x)f(y)P(x, dy)\pi(dx) \geq 0, \qquad \forall f \in L^2(\pi). \tag{9}$$

This is not the standard definition of positive operators; more comments on this issue will be given in Section 3.2. When referring to a general state space we mean a state spaces equipped with a countably generated σ-field, i.e. generated by a countable collection of subsets of the state space. On finite state spaces an irreducible chain is called *aperiodic* if for some i (and hence for all) the greater common divisor of $\{ t > 0 : P(X_t = i | X_0 = i) \}$ is equal to one. On general state spaces an m-cycle for an irreducible chain is a collection $\{E_0, \ldots, E_{m-1}\}$ of disjoint sets such that $P(x, E_j) = 1$ for $j = i+1 \mod m$ and all $x \in E_i$. The period of the chain is the largest m for which an m-cycle exists. The chain is aperiodic if $d = 1$.

3 Efficiency ordering and Peskun ordering

In classical statistics, estimates are compared in terms of their asymptotic relative efficiency, likewise here we will prefer a Markov chain if it produces estimates that are asymptotically more efficient on a sweep-by-sweep basis:

Definition 3.1.
If P and Q are Markov chains with stationary distribution π, then P is at least as efficient as Q, $P \succeq_E Q$, if

$$v(f,P) \leq v(f,Q), \qquad \forall f \in L_0^2(\pi). \tag{10}$$

In order to compare Markov chains in term of their efficiency it is useful to refer to the partial ordering introduced by Peskun [25] for discrete state spaces and extended by Tierney [30] to general state spaces.

Definition 3.2.
If P and Q are Markov chains on a measurable space with stationary distribution π, then P dominates Q off the diagonal, $P \succeq Q$, if for $\pi-$almost all x in the state space we have

$$P(x, B \setminus \{x\}) \geq Q(x, B \setminus \{x\})$$

for all measurable B.

For a better understanding of this definition let us restrict our attention to finite state spaces. In this setting, P dominates Q off the diagonal if each of the off-diagonal elements of P is greater than or equal to the corresponding off-diagonal elements in Q. This means that P has higher probability of moving around in the state space than Q and therefore the corresponding Markov chain will explore the space in a more efficient way (better mixing). Thus, we expect that the resulting MCMC estimates will be more precise than the ones obtained by averaging along a Markov chain generated via Q. This intuition is stated more rigorously in the next theorem by Tierney [30] which holds on general state spaces.

Theorem 3.1.
Let P and Q be reversible transition kernels with stationary distribution π. If $P \succeq Q$, then $P \succeq_E Q$.

The following theorem also appears in [30]:

Theorem 3.2.
If P and Q have stationary distribution π and $P \succeq Q$, then $Q-P$ is a positive operator on $L^2(\pi)$.

In the next section we show how the Peskun ordering, Definition 3.2, implies an ordering on the eigenvalues of the corresponding transition matrices in finite state spaces. We extend the result to general state spaces in Section 3.2.

3.1 Finite State Spaces

Let $\{\lambda_{0P}, \lambda_{1P}, \dots\}$ be the eigenvalues of P, arranged in decreasing order, and let $\{e_{0P}, e_{1P}, \dots\}$ be the corresponding normalized right eigenvectors, so that $Pe_{jP} = \lambda_{jP}e_{jP}$, $j = 0, 1, \dots$. For $P \in \mathcal{P}$ there is an eigenvalue equal to one, λ_{0P}, which is associated with the constant eigenvector. Since this is always the case let us restrict our attention to the eigenvalues associated with non-constant eigenvectors. Reversibility of a transition kernel ensures that the eigenvalues and eigenvectors are real.

Theorem 3.3.
For $P, Q \in \mathcal{D}$, if $Q - P \geq 0$, then $\lambda_{iP} \leq \lambda_{iQ}$ for all i.

Proof. Consider the following definition of eigenvalues [2]:

$$\lambda_{iP} = \min_{g_1,\dots,g_i} \max_{\substack{(f,g_j)=0 \\ j=1,\dots,i}} \frac{(f, Pf)}{(f, f)}$$

(the min is taken of all sets of vectors $g_1, \dots, g_i$). If $Q - P \geq 0$ then

$$\frac{(f, Qf)}{(f, f)} \geq \frac{(f, Pf)}{(f, f)}, \qquad \forall f \in L^2(\pi)$$

and the result follows since the eigenvalues of a transition matrix and of the corresponding operator in $L^2(\pi)$ are the same (because the defining equation is the same). □

The previous theorem is a known fact for symmetric matrices. In our setting neither P nor Q need to be symmetric but if we consider them as operators on $L^2(\pi)$ they are indeed self-adjoint operators, provided that the detailed balance condition holds.

By Theorem 3.2, $P \succeq Q$ implies that $Q - P \geq 0$, thus the Peskun ordering induces an ordering on all the eigenvalues of the two transition matrices. This proof can be generalized to compact operators on Hilbert spaces since their spectra are either empty, finite, or countable with zero as the only limit point [6]. But, as noticed in [5], not many Markov chains have compact transition operators.

Frigessi et al. [9] identify the subset of matrices in $\mathcal{P}$ which minimize $v(f, P)$ for all possible functions f. They begin by describing the structure of the matrices in $\mathcal{P}$ that have the smallest possible second largest eigenvalue. The procedure is then repeated in order to build a matrix with lowest third

eigenvalue, given that the second is already the smallest possible. By iterating, the matrix which is minimal with respect to the lexicographic order of the eigenvalues within $\mathcal{P}$ is obtained. This matrix gives rise to a Monte Carlo method with smaller asymptotic variance compared with independent sampling since all eigenvalues (except the largest one) turn out to be negative.

3.2 General State Spaces

In this section we extend the results obtained for finite state spaces to general state spaces. The difficulty lies in the fact that, while in finite state spaces we have a finite number of eigenvalues and it makes sense to compare and order eigenvalues of two transition matrices, in general state spaces we cannot talk about eigenvalues anymore but we need to introduce the concept of a spectrum. Let $\sigma(P)$ be the *spectrum* of P considered as an operator on $L^2(\pi)$, that is, the set of λ's such that $\lambda I - P$ is not invertible, where I denotes the identity operator on $L^2(\pi)$. The spectrum includes the eigenvalues, the λ's for which $\lambda I - P$ is not one-to-one. But it also includes the values λ such that $\lambda I - P$ is not onto. For linear operators on finite dimensional vector spaces, one-to-one and onto are equivalent so that $\sigma(P)$ is the set of the eigenvalues of P.

The *norm* of a linear operator on $L^2(\pi)$ is defined by

$$\|P\| = \sup_{\substack{u \in L^2(\pi) \\ u \neq 0}} \frac{\|Pu\|}{\|u\|}$$

where $\|u\|^2 = (u, u)$. The spectrum is a non-empty closed subset of the interval $[-1, +1]$ since the norm of P is less than or equal to one by Jensen's inequality and the norm of an operator bounds the spectrum (Proposition 1.11 (e) p. 239 in [6]). In this setting it does not make sense to say that the spectrum of one operator is smaller than the spectrum of another operator, we can at most compare the suprema of the spectra and this is what we will do. For reversible geometrically ergodic chains, all the eigenvalues but the principal eigenvalue, $\lambda_{0P} = 1$, are bounded away from ± 1 [27].

When considering a transition kernel as an operator on the subset $L_0^2(\pi)$ of $L^2(\pi)$ of zero mean functions, we eliminate from its spectrum the eigenvalue one associated with constant functions. Unless otherwise stated a transition kernel will be considered as an operator on $L_0^2(\pi)$.

Let $l_P = I - P$ be the *Laplacian operator* of the chain. An operator is invertible if it is one-to-one and onto. In our setting, the Laplacian l_P is invertible if it has a trivial null space when considered as an operator on $L_0^2(\pi)$

(one-to-one) and if its range is the entirety of $L_0^2(\pi)$ (onto). By the definition of spectrum l_P is invertible if and only if the spectrum of P does not contain the point 1. By the open mapping theorem ([6] Chapter 3, Theorem 12.5) invertible means that there exists an inverse which is a bounded operator on $L_0^2(\pi)$.

If a Markov chain has an invertible Laplacian, then the CLT (2) holds for the stationary chain for every function $f \in L_0^2(\pi)$ [12].

A weaker requirement on the Laplacian is that it is injective (one-to-one). In terms of the spectrum of P this is equivalent to the fact that one is not an eigenvalue. In this case the CLT holds for every function in the range of l_P [12]. For every such function we can still talk about the inverse Laplacian if we restrict the domain of l_P^{-1} to be the range of l_P. In other words, for any f in the range of l_P there exists a $g \in L_0^2(\pi)$ such that $f = l_P g$ so that $g = l_P^{-1} f$.

For a recurrent Markov chain, the only functions in $L^2(\pi)$ satisfying $Pf = f$ or equivalently $l_P f = 0$ (harmonic functions) are π-almost surely constant (Proposition 17.4.1 in [22] and [12]). Thus the operator P is injective and the Laplacian, as an operator on $L_0^2(\pi)$, has a trivial null space $= \{0\}$. Since our chains are recurrent, in our setting the Laplacian is an injective operator.

Let $E_P(\cdot)$ be the *resolution of the identity* associated with P in the spectral theorem [6], that is,

$$P = \int \lambda E_P(d\lambda).$$

As in [6], for every bounded Borel measurable function g on $\sigma(P)$ define

$$g(P) = \int g(\lambda) E_P(d\lambda).$$

In general our integrals will not involve the resolution of the identity but the spectral measure which is defined below. Given a function g in $L_0^2(\pi)$ define $E_{g,P}(\cdot) = (g, E_P(\cdot)g)$ to be the *spectral measure* associated with g. Then

$$(g, f(P)g) = \int f(\lambda) E_{g,P}(d\lambda)$$

for all bounded measurable functions f.

If P is irreducible, then $E_{g,P}$ is a positive measure on $[-1, +1)$ because atoms in the spectrum are eigenvalues (Proposition 12.29 (c) in [28]) and, as noted before, 1 is not an eigenvalue when considering P as an operator on $L_0^2(\pi)$. Let $\lambda_{\max,P} = \sup\{\, \lambda : \lambda \in \sigma(P) \,\}$ and $\Lambda_{\max,P} = \sup\{\, |\lambda| : \lambda \in \sigma(P) \,\}$. $\Lambda_{\max,P}$ is also called the spectral radius. The quantity $1 - \Lambda_{\max,P}$ is the spectral gap. If a transition kernel P has $1 - \Lambda_{\max,P} > 0$, we say that it has

a spectral gap. Roberts and Rosenthal in [27] show that a Markov chain is geometrically ergodic if and only if it has a spectral gap.

A reversible transition kernel P as an operator on $L_0^2(\pi)$ is a self-adjoint contraction, so the Laplacian is a positive operator and has a square root, $l_P^{\frac{1}{2}}$. If the chain is irreducible then $l_P^{\frac{1}{2}}$ is also injective and self-adjoint and therefore its range is dense and $l_P^{-\frac{1}{2}}$ is also self-adjoint ([6] p. 309). As proved in [17] the range of $l_P^{\frac{1}{2}}$ is indeed the set of functions that have a finite asymptotic variance. Another interesting result from [17] is that, for a stationary, irreducible, reversible transition kernel P, the variance of a function g in the CLT can be written as

$$v(g,P) = \int_{-1}^{1} \frac{1+\lambda}{1-\lambda} E_{g,P}(d\lambda). \tag{11}$$

Denote the *domain* and *range* of an operator A by $D(A)$ and $R(A)$, respectively. An operator on $L_0^2(\pi)$ is said to be *densely defined* if $D(A)$ is dense in $L_0^2(\pi)$. An operator is *positive*, $A \geq 0$, if $(g, Ag) \geq 0$, $\forall g \in L_0^2(\pi)$. Notice that, if we restrict ourselves to the space of real-valued functions in $L_0^2(\pi)$, then the fact that an operator is positive does not imply that the operator is self-adjoint. If, on the other hand, we consider also complex-valued functions, then $A \geq 0$ implies $A = A^*$, where A^* is the adjoint of A. There are functional analysis books such as [6], Theorem 3.8, that claim that the only positive operators are self-adjoint. This is because they are considering complex-valued spaces but this is not explicitly stated in the theorem (but 50 pages before). This fact can be quite misleading. In [20], the authors explicitly consider the space of complex valued functions. Since real valued functions are the only functions we are interested in, from a statistical point of view, we restrict ourselves to such functions when dealing with non-reversible Markov chains so that when we require a non-self-adjoint operator to be positive we do not contradict ourselves. The next lemma and corollary will be used in Section 4.

Lemma 3.1.
Let A be a positive, self-adjoint, injective, bounded operator. Then, for every $g \in D(A)$

$$(g, Ag) = \sup_{f \in D(A^{-\frac{1}{2}})} \left[2(f,g) - (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f)\right].$$

Proof. Since A is positive A^{-1} is also positive. This allows us to take square roots of both A and A^{-1}. Let $h = Ag$ so $g = A^{-1}h$. Clearly $D(A^{-1}) \subset$

$D(A^{-\frac{1}{2}})$ and for every $f \in D(A^{-\frac{1}{2}})$

$$\begin{aligned} 0 &\leq (A^{-\frac{1}{2}}(f-h), A^{-\frac{1}{2}}(f-h)) \\ &= (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f) - 2(A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}h) + (A^{-\frac{1}{2}}h, A^{-\frac{1}{2}}h). \end{aligned}$$

Now substitute $h = Ag$ and use the fact that $(f,g) = (g,f)$, which is true in a real Hilbert space but not true in complex Hilbert spaces. Thus

$$(g, Ag) \geq \left[2(f,g) - (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f)\right], \qquad \forall f \in D(A^{-\frac{1}{2}}) \tag{12}$$

and the supremum is achieved by taking $f = h$ since, in this case, the right hand side equals the left hand side in (12). □

Corollary 3.1.
Suppose A and B are positive, self-adjoint, injective, bounded operators. If the two conditions

$$(B^{-\frac{1}{2}}f, B^{-\frac{1}{2}}f) \leq (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f), \qquad \forall f \in D(A^{-\frac{1}{2}})$$

and

$$D(A^{-\frac{1}{2}}) \subset D(B^{-\frac{1}{2}})$$

are satisfied, then $A \leq B$.

Proof. By Lemma 3.1 we have for every $g \in D(A) = D(B)$

$$\begin{aligned} (g, Bg) &= \sup_{f \in D(B^{-\frac{1}{2}})} \left[2(f,g) - (B^{-\frac{1}{2}}f, B^{-\frac{1}{2}}f)\right] \\ &\geq \sup_{f \in D(A^{-\frac{1}{2}})} \left[2(f,g) - (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f)\right] \\ &= (g, Ag). \end{aligned}$$

□

In this section we will make extensive use of the following result.

Lemma 3.2.
For a transition kernel P with stationary distribution π, the asymptotic variance can be written as

$$v(g,P) = (g, [2l_P^{-1} - I]g), \qquad \forall g \in D(l_P^{-1}). \tag{13}$$

Proof. For any $g \in D(l_P^{-1})$ there exists an $f \in L_0^2(\pi)$ such that $g = l_P f$ so that $Pf = f - g$. Using a result in [12] we can write the asymptotic variance as

$$\begin{aligned} v(g,P) &= \|f\|^2 - \|Pf\|^2 \\ &= \|f\|^2 - \|f-g\|^2 \\ &= (f,f) - (f-g, f-g) \\ &= 2(g,f) - (g,g) \\ &= 2(g, l_P^{-1} g) - (g,g) \\ &= (g, [2l_P^{-1} - I]g). \end{aligned}$$

□

The previous result generalizes the representation of the asymptotic variance given in [16] for finite state spaces. Notice that the transition kernel does not need to be reversible for this lemma to hold.

The next theorem extends Theorem 3.3 to general state spaces.

Theorem 3.4.
Given reversible Markov chains P and Q with stationary distribution π, suppose $P \succeq Q$, then

$$\lambda_{\max,P} \leq \lambda_{\max,Q}. \tag{14}$$

Proof. It follows directly from Theorem X.4.2 of [8] that for any bounded self-adjoint operator A on a Hilbert space we have

$$\lambda_{\max,A} = \sup_{\|f\|=1} (f, Af).$$

Thus (14) holds whenever $Q - P \geq 0$, and Theorem 3.2 finishes the proof. □

3.3 A Counterexample

The rate of convergence in total variation distance of P^n (and of weak convergence of X_n) to $\pi(x)$ is governed by the spectral radius, $\Lambda_{\max,P}$, which, in finite state spaces is the second largest eigenvalue in absolute value [4, 11]. We thus have conflicting requirements: fast total variation convergence to equilibrium is obtained by having all eigenvalues small in absolute value while good properties in terms of asymptotic variance of ergodic averages are obtained

by having small positive and large negative eigenvalues, as (11) indicates. Only if the transition kernels are positive operators, that is, if the eigenvalues are all positive, are the two goals not in conflict. It has been shown by Liu et al. that the independence Metropolis-Hastings algorithm [19] and the random scan Gibbs sampler [20] are positive operators. Mira and Tierney in [23] prove that the slice sampler is also a positive operator.

The next example shows that the Peskun partial ordering does *not* imply an ordering on the largest eigenvalue in absolute value. Consider the following two transition matrices:

$$A = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

Matrix A refers to a symmetric random walk with reflecting barriers at the end points. With B, no matter where you are there is equal probability to move to any of the other 2 states. Both transition matrices, being doubly stochastic, have $\pi = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$ as their stationary distribution.

Clearly $B \succeq A$. The ordered eigenvalues of the two matrices are

$$\lambda_A = [1,\ 0.5,\ -0.5]$$

and

$$\lambda_B = [1,\ -0.5,\ -0.5].$$

As expected from Theorem 3.2,

$$\lambda_{iB} \leq \lambda_{iA}, \qquad \forall i.$$

Consider now the transition matrix given by $C = 0.8\,A + 0.1\,B + 0.1\,I_3$ where I_n is the identity matrix of dimension n. We have

$$C = \begin{pmatrix} 0.5 & 0.45 & 0.05 \\ 0.45 & 0.1 & 0.45 \\ 0.05 & 0.45 & 0.5 \end{pmatrix}$$

and $\lambda_C = [1,\ 0.45,\ -0.35]$. Again the stationary distribution of C is π, $B \succeq C$ and

$$\lambda_{iB} \leq \lambda_{iC}, \qquad \forall i$$

but $\Lambda_{\max,B} > \Lambda_{\max,C}$.

4 A new ordering

The Peskun criterion and its generalization given by Tierney order only a limited number of Markov chains. For example, the ordering does not allow comparing two distinct transition matrices having all zeros on the main diagonal or two transition kernels for which $P(x, \{x\}) = 0$ for every x in the state space. The latter includes all Gibbs samplers with continuous full conditional distributions. Furthermore, if only one of the off-diagonal entries of $P - Q$ is "out of order" then P and Q are incomparable.

A natural way to define a weaker ordering for comparing more Markov chains is the following.

Definition 4.1.
P dominates Q in the covariance ordering, $P \succeq_1 Q$, if $Q - P$ is a positive operator on $L_0^2(\pi)$, that is, if $(f, (Q-P)f) \geq 0$, for every $f \in L_0^2(\pi)$.

Restricting ourselves to $L_0^2(\pi)$ does not reduce the generality of the previous definition, since

$$(f, Qf) \geq (f, Pf), \qquad \forall f \in L_0^2(\pi)$$

if and only if

$$(f, Qf) \geq (f, Pf), \qquad \forall f \in L^2(\pi).$$

One implication is obvious. For the other, let f in $L^2(\pi)$, then $f_0 = f - \mu$ with $f_0 \in L_0^2(\pi)$ and $(f, Pf) = (f_0, Pf_0) + \mu^2$. Similarly we have $(f, Qf) = (f_0, Qf_0) + \mu^2$ and this gives what we want.

The binary relation $\succeq_1$ defines a partial ordering on the space $\mathcal{D}$ of reversible Markov chains with respect to π, since the following properties hold:

1. **Reflexive.** $P \succeq_1 P$ since $(f, (P-P)f) \geq 0$ for all $f \in L_0^2(\pi)$.

2. **Antisymmetric.** $P \succeq_1 Q$ and $Q \succeq_1 P$ imply $P = Q$. This means that $(f, (Q-P)f) \geq 0$ and $(f, (P-Q)f) \geq 0$ for all $f \in L^2(\pi)$ imply $P = Q$. In order to prove this, it is sufficient to show that $(f, Af) = 0$ for all $f \in L^2(\pi)$ implies that A is the zero operator. This in turn is equivalent to

$$(a+b, A(a+b)) = 0, \qquad \forall a, b \in L^2(\pi). \tag{15}$$

But since $A \in \mathcal{D}$

$$(a+b, A(a+b)) = (a, Aa) + (b, Ab) + 2(a, Ab).$$

By assumption the first two terms on the right hand side are equal to zero, thus condition (15) is equivalent to

$$(a, Ab) = 0, \qquad \forall a, b \in L^2(\pi)$$

which implies that A is the zero operator as required.

3. **Transitive.** $P \succeq_1 Q$ and $Q \succeq_1 R$ implies $P \succeq_1 R$. This is easy to verify since, if $(f, (Q-P)f) \geq 0$ and $(f, (R-Q)f) \geq 0$ for all $f \in L_0^2(\pi)$, then, $(f, (Q-P)f) + (f, (R-Q)f) = (f, (R-P)f) \geq 0$ for all $f \in L_0^2(\pi)$.

Notice that, if we consider also non-self-adjoint operators and move from $\mathcal{D}$ to $\mathcal{P}$, then $\succeq_1$ is not a partial ordering anymore since the antisymmetry property fails. To see this consider a non-reversible transition kernel P and let P^* be its adjoint. Then $(f, Pf) = (f, P^*f)$ so that $P \succeq_1 P^*$ and $P^* \succeq_1 P$ but it is not true that $P^* = P$ unless P is self-adjoint.

The condition $P \succeq_1 Q$ is equivalent to

$$\mathrm{Cov}_\pi(f, Qf) \geq \mathrm{Cov}_\pi(f, Pf), \qquad \forall f \in L_0^2(\pi)$$

where

$$\mathrm{Cov}_\pi(f, Qf) = E_\pi[f(X_0)f(X_1)] = \gamma_1$$

is the lag one autocovariance.

Covariance order does not imply Peskun order, as the next example shows, but Peskun order does imply covariance order (Theorem 3.2). Hence covariance order is a more general (weaker) criterion.

Consider the following matrices:

$$P = \begin{pmatrix} 0.3 & 0.3 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.3 & 0.3 \\ 0.2 & 0.2 & 0.3 & 0.3 \end{pmatrix}$$

and

$$A = x^T x = \begin{pmatrix} 0.1 & 0.1 & -0.1 & -0.1 \\ 0.1 & 0.1 & -0.1 & -0.1 \\ -0.1 & -0.1 & 0.1 & 0.1 \\ -0.1 & -0.1 & 0.1 & 0.1 \end{pmatrix}$$

where $x = [\ \sqrt{0.1},\ \sqrt{0.1},\ -\sqrt{0.1},\ -\sqrt{0.1}\ ]$. The matrix P is doubly stochastic therefore it is a transition matrix with uniform stationary distribution. The matrix A is positive definite. Let $Q = P + A$. Since the row and column sums of A are zero, Q is again a doubly stochastic matrix so that both P and Q have the same stationary distribution. We have that $Q - P = A \geq 0$ therefore $P \succeq_1 Q$, but it is not true that $P \succeq Q$ because this would imply that the matrix $Q - P$ has all negative off diagonal elements which is not true.

The ordering we have introduced is equivalent to Löwner partial ordering, $(\geq_L)$, on positive, bounded, linear operator on a Hilbert space [21], [3]. Löwner ordering is defined on positive operators therefore we need to consider the Laplacian of P, $l_P = I - P$, instead of P. Since $P \succeq I$ for every $P \in \mathcal{P}$, we have that $l_P \geq 0$.

Definition 4.2.
Given two positive, bounded, self-adjoint, linear operators on a Hilbert space, l_P, l_Q, we say that l_P dominates l_Q in the Löwner sense, $l_P \geq_L l_Q$, if $l_P - l_Q \geq 0$.

The following conditions are equivalent:

1. $P \succeq_1 Q$ i.e. $Q - P \geq 0$;
2. $l_P \geq_L l_Q$ i.e. $l_P - l_Q \geq 0$.

A variety of inequalities are obtainable, for any partial ordering, once the order-preserving functions are identified. For the Löwner ordering or better for a generalization of it that does not require the operators to be positive, the following theorem characterizes the class of order preserving functions [21], [3]. Let $f(x)$ be a bounded real-valued function of a real variable x defined in an interval I. Consider a bounded self-adjoint operators A in a Hilbert space H whose spectrum lies in the domain of $f(x)$. Then by $f(A)$ we mean the self-adjoint operator defined as $f(A) = \int f(\lambda) E_A(d\lambda)$.

Theorem 4.1.
A necessary and sufficient condition for a continuous real-valued function f

on (I_1, I_2) *to have the property that* $f(A) \leq f(B)$ *for all pairs of bounded, self-adjoint operators* A *and* B *with* $\sigma(A), \sigma(B) \subseteq (I_1, I_2)$ *and* $A \leq B$ *is that* f *is analytic in* (I_1, I_2)*, can be analytically continued into the whole upper half-plane, and represents there an analytic function with the property* $(\text{Im } f) \geq 0$ *for all* z *with* $(\text{Im } z) > 0$.

Further characterizations of such classes of functions can be found in [18]. A function that satisfies the conditions of Theorem 4.1 is

$$h(x) = \frac{ax + b}{cx + d} \qquad \text{with } ad - bc > 0$$

either in $x > -\frac{d}{c}$ or $x < -\frac{d}{c}$.

Take $a = b = d = 1$ and $c = -1$, then $ad - bc = 2 > 0$, and

$$h(x) = \frac{1 + x}{1 - x}$$

preserves the ordering for $x < 1$. Thus

$$P \succeq_1 Q \text{ if and only if } Q \geq P \text{ if and only if } \frac{I + Q}{I - Q} \geq \frac{I + P}{I - P}.$$

The covariance ordering is equivalent to the efficiency ordering as the next theorem states. This provides a characterization of the efficiency ordering.

Theorem 4.2.
Let P *and* Q *be reversible and irreducible transition kernels with stationary distribution* π*. Then* $P \succeq_E Q$ *if and only if* $P \succeq_1 Q$.

Proof. Let us consider two cases depending on whether the Laplacian is an invertible operator on $L_0^2(\pi)$.

Case (1) Suppose l_P is invertible. Let $h(l_P) = \frac{2}{l_P} - I = \frac{I+P}{I-P}$. Using Lemma 3.2, $P \succeq_E Q$ holds if and only if, for all $f \in L_0^2(\pi)$,

$$(f, h(l_P)f) \leq (f, h(l_Q)f) \tag{16}$$

which, by definition is equivalent to

$$h(l_P) \leq h(l_Q) \tag{17}$$

and by Theorem 4.1, this is true if and only if

$$Q - P \geq 0, \tag{18}$$

which is $P \succeq_1 Q$.

Case (2) If l_P is not invertible, we have to prove the equivalence of (17) and (18) without using Theorem 4.1 on any noninvertible operators.

First we prove $P \succeq_1 Q$ implies $P \succeq_E Q$. So assume $P \succeq_1 Q$, and let $K_{\epsilon P} = I-(1-\epsilon)P$ for $0 < \epsilon < 1$. $K_{\epsilon P}$ is invertible since its spectrum $\sigma(K_{\epsilon P}) \subseteq (\epsilon, 2-\epsilon)$ does not contain zero. Furthermore $h(K_{\epsilon P})$ is also invertible since its spectrum is

$$\sigma(h(K_{\epsilon P})) = h(\sigma(K_{\epsilon P})) \subseteq \left(\frac{\epsilon}{2-\epsilon}, \frac{2-\epsilon}{\epsilon}\right).$$

Then, for all $0 < \epsilon < 1$, $Q - P \geq 0$ implies $K_{\epsilon Q} \leq K_{\epsilon P}$ and from case (1) this is true if and only if

$$(f, h(K_{\epsilon,Q})f) \geq (f, h(K_{\epsilon,P})f), \qquad \forall f \in L_0^2(\pi). \tag{19}$$

We now want to take the limit as $\epsilon \to 0$. Consider

$$(f, h(K_{\epsilon,P})f) = \int \frac{1+(1-\epsilon)\lambda}{1-(1-\epsilon)\lambda} E_{fP}(d\lambda)$$

The derivative of the integrand with respect to ϵ is

$$\frac{-2\lambda}{[1-(1-\epsilon)\lambda]^2}$$

thus, for $\lambda \in [-1, 0)$ the integrand is increasing in ϵ while for $\lambda \in [0, +1)$ the integrand is decreasing. This suggests that we break the integral over these two subsets of the spectrum

$$(f, h[K_{\epsilon,P}]f) = \int_{-1}^{0} \frac{1+(1-\epsilon)\lambda}{1-(1-\epsilon)\lambda} E_{fP}(d\lambda) + \int_{0}^{1} \frac{1+(1-\epsilon)\lambda}{1-(1-\epsilon)\lambda} E_{fP}(d\lambda)$$

For every $\lambda \in \sigma(P)$ and every $\epsilon \in (0, 1)$ the integrals are finite by construction, therefore a modified version of the standard monotone convergence theorem ([10] p. 50) can be used to take the limit inside the integral and we get that (19) implies (17). Hence $P \succeq_1 Q$ implies $P \succeq_E Q$.

Now we prove the implication in the other direction: $P \succeq_E Q$ implies $P \succeq_1 Q$. So assume $P \succeq_E Q$ so (17) holds. Then from the properties of the Laplacian recalled in Section 3.2 and in particular the fact that the range of $l_Q^{\frac{1}{2}}$ is the set of functions that have a finite asymptotic variance [17]

$$v(f, P) \leq v(f, Q) < \infty, \qquad \forall f \in R(l_Q^{\frac{1}{2}})$$

and

$$R(l_Q^{\frac{1}{2}}) \subseteq R(l_P^{\frac{1}{2}}).$$

It follows that

$$(l_P^{-\frac{1}{2}} f, l_P^{-\frac{1}{2}} f) \leq (l_Q^{-\frac{1}{2}} f, l_Q^{-\frac{1}{2}} f), \qquad \forall f \in R(l_Q^{\frac{1}{2}}) = D(l_Q^{-\frac{1}{2}}) \tag{20}$$

that is, and since the hypotheses of Corollary 3.1 are satisfied we have $l_Q \leq l_P$, hence $P \succeq_1 Q$. □

One application of Theorem 4.2 is given in the following corollary. If we have two transition kernels P and Q having the same stationary distribution there are different possible strategies to run our Markov chain. We could choose one of the two transition kernels and iterate it, obtaining P^n or Q^n respectively. Otherwise we could combine the two basic steps via composition, obtaining a hybrid sampler. If we know that one of the two original kernels is more efficient than the other, then the next corollary gives guidelines on how to combine to two kernels in an efficient way.

Corollary 4.1.
If $P \succeq_E Q$ then $P^3 \succeq_E PQP$ and $QPQ \succeq_E Q^3$.

Proof. The first inequality follows from

$$I - P \geq I - Q$$

by multiplying on both sides by Q. The second inequality follows by multiplying both sides by P. □

Another interesting theorem related to Löwner ordering is the following [13].

Theorem 4.3.
If $A \geq 0$ and $B \geq 0$ are Hermitian matrices, then $A \leq B$ if and only if $R(A) \subseteq R(B)$ and $\lambda_{\max}(AB^+) \leq 1$.

Here B^+ is any generalized inverse (not necessarily the Moore-Penrose generalized inverse). If P and Q are reversible transition kernels with respect to π, then $A = l_Q$ and $B = l_P$, are positive self-adjoint operators and

$$P \succeq_1 Q$$

$$\Updownarrow$$

$$v(f, P) \leq v(f, Q), \qquad \forall f \in L_0^2(\pi)$$

$$\Updownarrow$$
$$l_P^{-1} \leq l_Q^{-1}$$
$$\Updownarrow$$
$$l_Q \leq l_P$$
$$\Updownarrow$$
$$l_Q^{\frac{1}{2}} \leq l_P^{\frac{1}{2}}$$

These implications follow from the fact that the function $h(x) = x^a$ preserves the Löwner ordering when $0 \leq a \leq 1$ while the function $h(x) = \frac{1}{x}$ reverses the ordering [3]. By Theorem 4.3 it follows that $P \succeq_E Q$ implies

$$R(l_P^{\frac{1}{2}}) \subseteq R(l_Q^{\frac{1}{2}}).$$

From [17] recall that if a function f is in the range of $l_P^{\frac{1}{2}}$ then the MCMC estimate of $E_\pi[f(X)]$ obtained using P as the transition kernel, has finite asymptotic variance in the CLT. The previous chain of equivalence relations tells us that if the estimate of a function has finite asymptotic variance under P, than it also has finite asymptotic variance under Q whenever P is more efficient than Q.

We finally report another result related to Löwner ordering. We have not made much use of it but we believe it could lead to interesting results when comparing transition matrices in terms of the covariance ordering. In [1] the authors characterize the class of functions of more than one variable that preserve Löwner ordering. That is functions f such that $A \leq A_1$ and $B \leq B_1$ imply

$$f(A, B) + f(A_1, B_1) \geq f(A, B_1) + f(A_1, B)$$

where the matrices involved in the comparison are Hermitian matrices. Functions such that

$$f(p(A, B) + (1 - p)(A_1, B_1)) \leq pf(A, B) + (1 - p)f(A_1, B_1)$$

for $p \in [0, 1]$ are also characterized in the same paper.

5 Non-reversible Markov chains

Reversibility of a transition kernel with respect to π implies that π is the stationary distribution of the corresponding Markov chain, but reversibility is a much stronger condition than (3). While (3) places restrictions only on the

marginal distribution of X_t, (7) places restrictions on the joint distribution of (X_t, X_{t+1}) by requiring that, when X_t has the distribution π, then (X_t, X_{t+1}) has the same joint distribution as (X_{t+1}, X_t).

Reversibility is not necessary for MCMC, only having the correct stationary distribution. However, reversibility of a transition kernel ensures that the corresponding operator on $L_0^2(\pi)$ is self-adjoint and this is a very appealing property when studying the behavior of our Markov chain. We can use spectral theory and this makes the analysis much easier. Moreover the only simple way to show that an update mechanism has a specified stationary distribution is to show that it is reversible with respect to that distribution. However, it is a very common practice to construct a Markov chain for Monte Carlo that is non-reversible by combining reversible elementary update steps by composition. If P and Q are reversible and have the same stationary distribution, then PQ also has the same stationary distribution but is reversible only if P and Q are commuting operators, which very rarely holds. Recently there has been a growing interest in non-reversible Markov chains since [15], [24] and [7] constructed non-reversible Markov chains and showed that they have better properties in terms of convergence to stationarity in total variation distance than other reversible operators.

In this section we restrict our attention to transition kernels P that are not self-adjoint but for which l_P is invertible. One important fact is that Lemma 3.2 still holds in this setting and thus we have

Corollary 5.1.
Let P and Q be irreducible Markov chains with stationary distribution π such that both l_P and l_Q are invertible. Then $P \succeq_E Q$ if and only if $l_P^{-1} \leq l_Q^{-1}$.

Proof. The proof follows directly from the representation of the asymptotic variance in the CLT that appears in equation (13). □

Lemma 5.1.
Let A be a injective positive linear operator defined on a subspace $V = D(A)$ with inverse A^{-1} defined on $R(A) = D(A^{-1})$. For every $g \in D(A^{-1})$

$$(g, A^{-1}g) = \sup_{f \in D(A)} \left[(f, g) + (Af, A^{-1}g) - (f, Af)\right].$$

Proof. Since $g \in D(A^{-1})$, there exists $h = A^{-1}g \in D(A)$. Then, for every $f \in D(A)$,

$$\begin{aligned} 0 &\leq (f - h, A(f - h)) \\ &= (f, Af) - (f, Ah) - (h, Af) + (h, Ah). \end{aligned}$$

It follows, substituting $g = Ah$, that

$$(g, A^{-1}g) \geq [(f, g) + (Af, A^{-1}g) - (f, Af)], \qquad \forall f \in D(A)$$

and the supremum is achieved by taking $f = h$. □

Corollary 5.2.
Let A and B be positive and invertible operators such that

$$B^*B^{-1} = A^*A^{-1}. \tag{21}$$

Then $A - B \geq 0$ implies $B^{-1} - A^{-1} \geq 0$.

Proof. Apply Lemma 5.1 with A replaced by B. The condition $B^*B^{-1} = A^*A^{-1}$ is needed so that $(Bf, B^{-1}g) = (Af, A^{-1}g)$ for all f and g. □

Notice that $B^*B^{-1} = A^*A^{-1}$ or, equivalently, $(BA^{-1})^* = A^{-1}B$, automatically holds if A and B are self-adjoint. Moreover if (21) holds, either both A and B are self-adjoint or neither is.

Corollary 5.3.
Let A and B be positive and invertible operators such that (21) *holds, then $B^{-1} - A^{-1} \geq 0$ implies $A - B \geq 0$.*

Proof. Apply Lemma 5.1 with A replaced by A^{-1}. □

Theorem 5.1.
Let P and Q be irreducible transition kernels with stationary distribution π such that both l_P and l_Q are invertible. Assume that $(l_Q)^*l_Q^{-1} = (l_P)^*l_P^{-1}$. Then

$$P \succeq_E Q \quad \textit{if and only if} \quad P \succeq_1 Q$$

This theorem is the equivalent of Theorem 4.2 for non-reversible Markov chains. The price paid for non-reversibility is the extra requirement $(l_Q)^*l_Q^{-1} = (l_P)^*l_P^{-1}$.

Proof. From Corollary 5.1 $v(f, P) \leq v(f, Q)$ for every $f \in L_0^2(\pi)$ if and only if $l_P^{-1} \leq l_Q^{-1}$. By Corollary 5.2 and 5.3, if $(l_Q)^*l_Q^{-1} = (l_P)^*l_P^{-1}$ this is equivalent to $P \leq Q$. □

Let $P^\diamond = \frac{P+P^*}{2}$, where P^* is the adjoint of P. Then, for all $f \in L^2(\pi)$

$$\begin{aligned}(f, P^\diamond f) &= \left(f, \frac{P + P^*}{2} f\right) \\ &= \frac{1}{2}[(f, Pf) + (f, P^*f)] \\ &= \frac{1}{2}[(f, Pf) + (Pf, f)] \\ &= (f, Pf).\end{aligned} \tag{22}$$

This says that every statement about (f, Pf) is actually is a statement involving only the self-adjoint part $P^\diamond$ of P.

The following example shows that the implication

$$l_P \geq l_Q \rightarrow l_Q^{-1} \geq l_P^{-1} \tag{23}$$

does not hold in general. Consider the following transition matrices:

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$P^* = P^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$P^\diamond = \frac{P + P^*}{2} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

All transition matrices being doubly stochastic have the uniform distribution as their stationary distribution. The state space can be viewed as a circle, with states labeled from 0 to $n-1$ and the n-th state coincides with the origin, zero. The given matrices refer to the case where $n = 5$. Both P and P^* are non-symmetric (non-self-adjoint) and represent deterministic walks along the circle in anti-clockwise and clockwise directions respectively. They produce estimates with zero asymptotic variance for every function in $L_0^2(\pi)$. On the other hand, $P^\diamond$ represents a symmetric random walk on the circle. By letting n increase we can make the asymptotic variance of ergodic averages obtained using a symmetric random walk as large as we wish. This means that:

$$(f, l_P^{-1} f) < (f, l_{P^\diamond}^{-1} f), \qquad \forall f \in L^2(\pi).$$

Because of (22) we also have

$$(f, l_P f) = (f, l_{P^\diamond} f), \qquad \forall f \in L^2(\pi)$$

therefore (23) with $Q = P^\diamond$ does not hold. Notice that condition (21) is not verified in this setting since $P^\diamond$ is self-adjoint while P is not.

6 Constructing the inverse

In this section we consider Markov chains on finite state spaces and explain how to construct a generalized inverse that represents the inverse Laplacian. If there are n states, then $L^2(\pi)$ has dimension n and $L_0^2(\pi)$ has dimension $n-1$. If the Markov chain is irreducible, then the Laplacian $l_P = I - P$ is an invertible operator on $L_0^2(\pi)$, but it is never an invertible operator on $L^2(\pi)$.

The space $L_0^2(\pi)$ is a bit hard to work with resulting in messy formulas. So we seek a generalized inverse l_P^- that agrees with the inverse on $L_0^2(\pi)$, that is, $l_P^- l_P f_0 = f_0$ and $l_P l_P^- f_0 = f_0$ for every $f_0 \in L_0^2(\pi)$.

The space $L^2(\pi)$ is also a bit hard to work with because of its unusual inner product. Let S denote the state space, then the inner product is defined by

$$(f, g) = \sum_{x \in S} f(x) g(x) \pi(x), \qquad f, g \in L^2(\pi). \tag{24}$$

An operator $A : L^2(\pi) \to L^2(\pi)$ has a matrix representation $A(x, y)$ where x and y range over the state space S. The action of the operator is represented by the matrix multiplication

$$(Ag)(x) = \sum_{y \in S} A(x, y) g(y), \qquad x \in S.$$

The adjoint of A has the matrix representation

$$A^*(x, y) = \frac{\pi(x) A(x, y)}{\pi(y)}, \qquad x, y \in S. \tag{25}$$

Both (24) and (25) run afoul of our basic intuitions about linear algebra which are limited to thinking of all finite-dimensional vector spaces as being $\mathbb{R}^S$ for some finite set S and having inner product and adjoint defined by (24) and (25) with $\pi(x) = 1$ for all x. Thus we can also denote $\mathbb{R}^S$ as $L^2(\nu)$ where ν is counting measure on S. Because we understand $L^2(\nu)$ better than $L^2(\pi)$, we want to study the connections between them.

Assuming $\pi(x) > 0$ for all x, which follows from irreducibility, we define

$$\tilde{P}(x, y) = \frac{\sqrt{\pi(x)} P(x, y)}{\sqrt{\pi(y)}}. \tag{26}$$

If P is a self-adjoint operator on $L^2(\pi)$, that is, if the detailed balance condition holds, then $\tilde{P}(x, y)$ is a self-adjoint operator on $L^2(\nu)$, and vice versa. Define another linear map T by

$$(Tf)(x) = \sqrt{\pi(x)} f(x). \tag{27}$$

Its matrix representation is the diagonal matrix with elements $\sqrt{\pi(x)}$. This is invertible with inverse defined by

$$(T^{-1}f)(x) = \frac{f(x)}{\sqrt{\pi(x)}}. \tag{28}$$

We can now rewrite $\tilde{P}$ as

$$\tilde{P} = TPT^{-1}. \tag{29}$$

Another way to indicate (29) is by the commutative diagram

$$\begin{array}{ccc} L^2(\pi) & \xrightarrow{P} & L^2(\pi) \\ {\scriptstyle T^{-1}}\uparrow & & \downarrow{\scriptstyle T} \\ E & \xrightarrow{\tilde{P}} & F \end{array}$$

where we write E and F for the Hilbert spaces that are the domain and codomain of $\tilde{P}$. They are, of course, finite-dimensional vector spaces, the question is what inner product they have. The diagram shows that

$$E \xrightarrow{T^{-1}} L^2(\pi) \xrightarrow{T} F$$

thus E and F are the same Hilbert space. Denote the inner product on E by $(\cdot,\cdot)_E$. Then, considering T a Hilbert space isomorphism,

$$(f,g)_E := (T^{-1}f, T^{-1}g) = \sum_{x\in S} \frac{f(x)}{\sqrt{\pi(x)}} \cdot \frac{g(x)}{\sqrt{\pi(x)}} \cdot \pi(x),$$

which is the usual inner product on $L^2(\nu)$. Hence E "is" $L^2(\nu)$. Thus our commutative diagram becomes

$$\begin{array}{ccc} L^2(\pi) & \xrightarrow{P} & L^2(\pi) \\ {\scriptstyle T^{-1}}\uparrow & & \downarrow{\scriptstyle T} \\ L^2(\nu) & \xrightarrow{\tilde{P}} & L^2(\nu) \end{array}$$

Let us now see what properties $\tilde{P}$ inherits from P. Since P is a stochastic matrix we have that $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ denotes the column vector of ones. Thus

$$\tilde{P}\sqrt{\pi} = TPT^{-1}\sqrt{\pi} = TP\mathbf{1} = T\mathbf{1} = \sqrt{\pi}. \tag{30}$$

Similarly, $\pi^T P = \pi^T$, since π is the stationary distribution for P. It follows that

$$\sqrt{\pi^T}\tilde{P} = \sqrt{\pi}^T T P T^{-1} = \pi^T P T^{-1} = \pi^T T^{-1} = \sqrt{\pi^T}. \tag{31}$$

Thus, $\tilde{P}$ has the same right and left eigenvectors, namely $\sqrt{\pi}$, corresponding to the eigenvalue 1. Notice that $\sqrt{\pi} = T\mathbf{1}$, so $T : L^2(\pi) \to L^2(\nu)$ maps the constants to multiples of $\sqrt{\pi}$. This suggests that, in order to construct the inverse of l_P, we perform a singular value decomposition on $l_{\tilde{P}} = I - \tilde{P} = UDV^T$. Then $l_{\tilde{P}}^- = VD^-U^T$ is the Moore-Penrose generalized inverse of $l_{\tilde{P}}$ [26]. From (30) it follows that $l_{\tilde{P}}\sqrt{\pi} = 0$, thus $DV^T\sqrt{\pi} = 0$. Similarly, from (31), $\sqrt{\pi}l_{\tilde{P}} = 0$, thus $DU^T\sqrt{\pi} = 0$. If we denote by V_j and U_j the j^{th} column of V and U respectively, we have that, for all j

$$d_{jj}(V_j, \sqrt{\pi}) = 0$$

and

$$d_{jj}(U_j, \sqrt{\pi}) = 0$$

thus, for the j^* such that $d_{j^*j^*} \neq 0$,

$$(V_{j^*}, \sqrt{\pi}) = 0$$

and

$$(V_{j^*}, \sqrt{\pi}) = 0.$$

There are $n-1$ non-zero $d_{j^*j^*}$, and the corresponding collection of V_{j^*} spans the subspace of $L^2(\nu)$ orthogonal to $\sqrt{\pi}$. A similar reasoning holds for U_{j^*}. Using the maps T and T^{-1} we now move everything back to $L^2(\pi)$

$$l_P^- = T^{-1}VD^-U^TT \tag{32}$$

and

$$l_P = T^{-1}UDV^TT. \tag{33}$$

Some of the properties of the operators defined in (32) are studied in the sequel. First notice that, since $UU^T = VV^T = I$, l_P^- and l_P commute, that is,

$$l_P^- l_P = T^{-1}VD^-DV^TT = T^{-1}UD^-DU^TT = l_P l_P^-.$$

From $T\mathbf{1} = \sqrt{\pi}$ and $DV^T\sqrt{\pi} = 0$ it follows that l_P, $l_P^-l_P$ (and hence also $l_Pl_P^-$) annihilate constant vectors. Furthermore, for every $f_0 \in L_0^2(\pi)$, $l_P^-l_Pf_0 = f_0$ and $l_Pl_P^-f_0 = f_0$ as requested. This is easy to verify if we pick the columns of V as a basis for $L^2(\nu)$. One column is $\sqrt{(\pi)}$ and it is the image of $\mathbf{1}$ under T. Take the images under T of the other columns as a basis for L_0^2. Now, $l_P^-l_Pf_0 = T^{-1}VD^-DV^TTf_0$, and D^-D is a diagonal matrix with all ones except for a zero on the main diagonal. Furthermore $VV^T = T^{-1}T = I$, and the result follows.

Consider finally the operator $I - l_P^-l_P$, that, with an excess of notation, we could define to be $l_{l_P^-l_P}$. For every $f \in L^2(\pi)$, $l_{l_P^-l_P}f = E_\pi(f) = \sum_x f(x)\pi(x)$. This follows from the the fact that a generic element f in $L^2(\pi)$ can be written as $f = f_0 + \pi^T f$ where $f_0 \in L_0^2(\pi)$, and $\pi^T f = \mu$ is the mean of f with respect to π. Thus

$$l_{l_P^-l_P}f = (I - l_P^-l_P)f = f - (l_P^-l_P)(f_0 + \pi^T f) = f - (l_P^-l_P)f_0 = f - f_0 = \pi^T f.$$

7 Examples

In this section we analyze some non-reversible Markov chains by means of the tools we have developed here. The first example is the same one studied in [7]. Consider a finite state space, $\{1, 2, \ldots, n\}$, with the uniform distribution, $\pi(x) = \frac{1}{n}$ as the target distribution. The nearest-neighbor symmetric random walk with holding probabilities of $\frac{1}{2}$ at each end is a reversible Markov chain converging to π. In order to avoid the diffusive behavior of the random walk, Diaconis et al. [7] propose to enlarge the state space by introducing an additional copy of each state. We relabel state s as $(+, s)$ and label its copy $(-, s)$, for $s = 1, \ldots, n$. The transition matrix considered in [7] switches between copies at rate $\frac{c}{n}$ for some value of $0 \leq c \leq n$. The other possible moves allowed are to the left and they happen with probability $1 - \frac{c}{n}$. This Markov chain has $\frac{\pi(x)}{2}$ as its stationary distribution on each half of the enlarged state space and hence the marginal distribution on the second component of the state (ignoring the $+$ or $-$ sign) is $\pi(x)$, as required. Notice that, since the stationary distribution is uniform, it does not really matter which two states we collapse to go back to the original state. Any sort of grouping of the states two by two preserves the stationary distribution on the original state space.

Let us relabel the state space in the following way: $(+, s) = s$, and $(-, s) = -s$. For $n = 3$ the Markov chain we study can be represented as in Figure (1) and the corresponding transition matrix on the enlarged state space is

$$P_c = \begin{pmatrix} 0 & 1-\frac{c}{3} & 0 & 0 & \frac{c}{3} & 0 \\ 0 & 0 & 1-\frac{c}{3} & \frac{c}{3} & 0 & 0 \\ 0 & 0 & \frac{c}{3} & 1-\frac{c}{3} & 0 & 0 \\ 0 & \frac{c}{3} & 0 & 0 & 1-\frac{c}{3} & 0 \\ \frac{c}{3} & 0 & 0 & 0 & 0 & 1-\frac{c}{3} \\ 1-\frac{c}{3} & 0 & 0 & 0 & 0 & \frac{c}{3} \end{pmatrix} \tag{34}$$

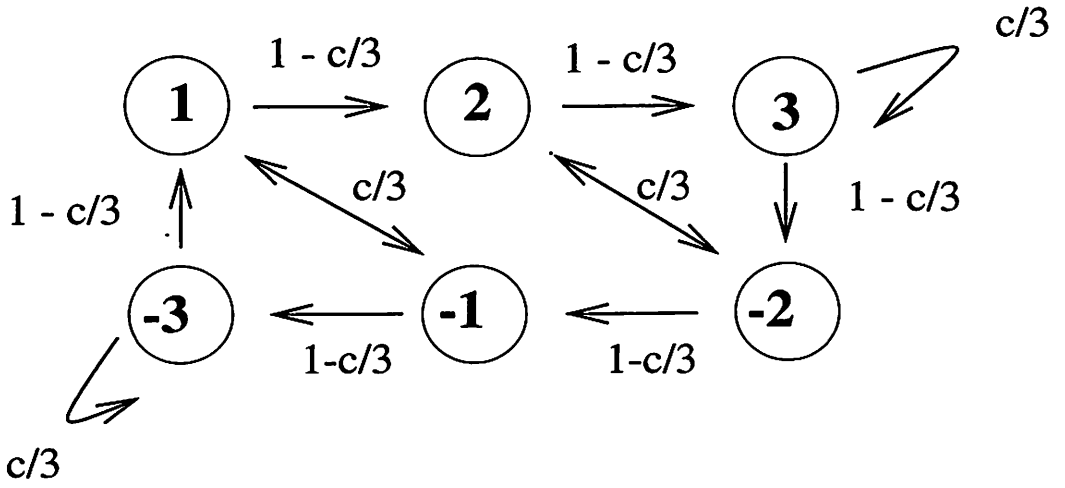

Figure 1: Enlarged state space

The rows and the columns of the matrix are labeled as in Figure (1) starting from state 1 and proceeding clockwise all the way up to state -3. We can think of the state space as a circle and the Markov chain either moves around the circle, with probability $1 - \frac{c}{n}$, or jumps across the circle, with probability $\frac{c}{n}$.

The operator corresponding to P_c is not self-adjoint because P_c does not satisfy the detailed balance condition. Since the stationary distribution is uniform, an operator on $L^2(\pi)$ is self-adjoint if and only if its transition matrix is symmetric.

By taking the inverse of l_{P_c} as described in Section 6, and letting the value of c vary over the interval $[0, n]$ we can study the properties of P_c in terms of asymptotic variance of the corresponding MCMC estimates. Using

Mathematica 3.0 we find that the eigenvalues of $(I - P_c)^- - (I - P_{c'})^-$ are

$$\begin{aligned}
\lambda_0 &= \lambda_1 = \lambda_2 = 0 \\
\lambda_3 &= \frac{6(c - c')}{(c' - 3)(c - 3)} \\
\lambda_4 &= \frac{2(c - c')}{(c' - 3)(c - 3)} \\
\lambda_5 &= \frac{3(c - c')}{2(c' - 3)(c - 3)}
\end{aligned}$$

Since $c \leq 3$ and $c' \leq 3$ these eigenvalues are non-negative if $c \geq c'$. This means that $v(f, P_c) \geq v(f, P_{c'})$ for every $f \in L_0^2(\pi)$ if $c \geq c'$. In other words, the performance of the transition matrix P_c in terms of asymptotic variance of *any function* of interest improves as c decreases towards 0, that is, as the probability of moving around the circle increases while the probability of jumping across the circle decreases.

The inverse Laplacian, as a function of the parameter c, is

$$l_{P_c}^- = \frac{1}{36(c-3)} \times \begin{pmatrix} -45+4c & -27+16c & -9+16c & 9+4c & 27-20c & 45+20c \\ 45-8c & -45+4c & -27+4c & -9-8c & 9+4c & 27+4c \\ 27+4c & 45+20c & -45-20c & -27+4c & -9+16c & 9+16c \\ 9+4c & 27-20c & 45+20c & -45+4c & -27+16c & -9+16c \\ -9-8c & 9+4c & 27+4c & 45-8c & -45+4c & -27+4c \\ -27+4c & -9+16c & 9+16c & 27+4c & 45+20c & -45-20c \end{pmatrix} \tag{35}$$

Let $c = 0$ and compute $2(I - P_c)^- - I$, which is the quantity that matters when computing the asymptotic variance of MCMC estimates

$$2l_{P_0}^- - I = \begin{pmatrix} -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} \\ -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} \end{pmatrix}$$

The previous matrix is not self-adjoint. Its self-adjoint part is

$$[(2l_{P_0}^- - I) + (2l_{P_0}^- - I)] = -\frac{1}{6}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Thus, the asymptotic variance of any function which is square integrable with respect to π, and has zero mean is zero. For this choice of the parameter the Markov chain on the enlarged state space moves around the state space in a deterministic fashion, that is, it circles around clockwise.

The eigenvalues of P_c are

$$\begin{aligned} \lambda_0 &= 1 \\ \lambda_1 &= \frac{2c-3}{3} \\ \lambda_2 &= \frac{3-c-\sqrt{-27+18c+c^2}}{6} \\ \lambda_3 &= \frac{-3+c-\sqrt{-27+18c+c^2}}{6} \\ \lambda_4 &= \frac{3-c+\sqrt{-27+18c+c^2}}{6} \\ \lambda_5 &= \frac{-3+c+\sqrt{-27+18c+c^2}}{6}. \end{aligned}$$

Setting $c = 0$ we get that the only real eigenvalues are $+1$ or -1 thus this transition matrix gives rise to a periodic Markov chain that does not converge to stationarity in total variation distance but produces MCMC estimates with zero asymptotic variance for any function of interest.

Diaconis et al. [7] compute the optimal value of c in terms of convergence to stationarity in χ^2 distance. Roughly this is $c = \sqrt{\log\, n}$ where n is the number of states for the original problem. This is an example of conflicting goals: the most efficient transition matrix differs from the one which is optimal in terms of speed of convergence to stationarity.

8 Comparing the performance of reversible and non-reversible kernels

Since reversible Markov chains are much easier to analyze it would be nice to have a way of transforming a non-reversible kernel, say P, into a reversible one, say Q, that is just as efficient. Recall that the performance of P in terms of asymptotic variance is related to $(I-P)^{-}$. Furthermore, as we have commented earlier, an operator P and its adjoint have the same asymptotic variance and so does $\frac{P+P^*}{2}$. Therefore $l_Q^{-1} = \frac{(I-P)^{-1}+[(I-P)^{-1}]^*}{2}$ is the inverse Laplacian of an operator with the same performance as P and to obtain Q from l_Q^{-1} we take the inverse and subtract what we get from the identity operator:

$$Q = I - \left\{ \frac{(I-P)^{-1} + [(I-P)^{-1}]^*}{2} \right\}^{-1} \tag{36}$$

$$= I - \left\{ \frac{(I-P)^{-1} + (I-P^*)^{-1}}{2} \right\}^{-1}. \tag{37}$$

Thus we only need to check if Q is a self-adjoint transition kernel with the proper stationary distribution. Notice that if P is self-adjoint then $Q = P$ as it would be sensible to require.

Q is self-adjoint since the identity is self-adjoint with respect to any stationary distribution, furthermore, for any operator A, the operator $(\frac{A+A^*}{2})^{-1}$ is self-adjoint with respect to the stationary distribution of A and the sum of self-adjoint operators is self-adjoint.

Q is as efficient as P since, as noted before,

$$(f, (I-Q)^{-1}f) = (f, \frac{(I-P)^{-1} + (I-P^*)^{-1}}{2} f) = (f, (I-P)^{-1}f),$$

for all $f \in L_0^2(\pi)$.

Let us now focus on the requirement that Q is a Markov transition kernel with the proper stationary distribution. We need to verify that

1. $\pi Q = \pi$;
2. $Q1 = 1$;
3. $f \geq 0$ implies $Qf \geq 0$;

where 1 represents the constant unitary function. For a transition operator A, the condition $\pi A = \pi$ holds if and only if $\pi Af = \pi f$ for all f in $L^2(\pi)$.

Using inner product notation, this means $(1, Af) = (1, f)$, for all f in $L^2(\pi)$. By the definition of adjoint, this is equivalent to $(A^*1, f) = (1, f)$, for all f in $L^2(\pi)$, which implies $A^*1 = 1$ by the Riesz representation theorem. Thus the condition $\pi A = \pi$ is equivalent to $A^*1 = 1$. Notice that the matrix condition $\pi^T A = \pi^T$ and $A\mathbf{1} = \mathbf{1}$ do not represent restrictions on A and A^* as operators on $L_0^2(\pi)$, though they determine the fact that these operators map into $L_0^2(\pi)$. This is easy to see since $A1 = 1$ is equivalent to $l_A 1 = 0$ or $l_A : L^2(\pi) \to L_0^2(\pi)$. Similarly $A^*1 = 1$ is equivalent to $l_{A^*} : L^2(\pi) \to L_0^2(\pi)$. If we start with the operator l_A defined only on $L_0^2(\pi)$ there is a unique extension to $L^2(\pi)$ that behaves like the Laplacian of a Markov transition operator, that is, satisfies $l_P 1 = 0$. A generic function $f \in L^2(\pi)$ can be written as $f_0 + c1$ where $f_0 \in L_0^2(\pi)$ and $c \in \mathbb{R}$. By linearity $l_A f = l_A f_0$, and similarly $l_{A^*} f = l_{A^*} f_0$. Hence the only thing to be shown about Q in order to verify requirements (1) and (2) is that $Q : L_0^2(\pi) \to L_0^2(\pi)$, which is true by definition, and we then extend Q and Q^* to $L^2(\pi)$ so that $Q1 = 1$ and $Q^*1 = 1$. In more detail, we have $l_P : L_0^2(\pi) \to L_0^2(\pi)$ and, assuming l_P is invertible, $l_P^{-1} : L_0^2(\pi) \to L_0^2(\pi)$. This implies $(l_{P^*})^{-1} = (l_P^{-1})^* : L_0^2(\pi) \to L_0^2(\pi)$, hence $\frac{1}{2}(l_P^{-1} + (l_{P^*})^{-1}) : L_0^2(\pi) \to L_0^2(\pi)$, again assume invertibility and call l_Q the inverse. It follows that $l_Q : L_0^2(\pi) \to L_0^2(\pi)$. Extend now l_Q from $L_0^2(\pi)$ to $L^2(\pi)$ by "reconstruction":

$$l_Q f = l_Q[f - (1, f)1], \qquad \forall f \in L^2(\pi)$$

where $(1, f)1$ is nothing but the mean of f under the stationary distribution. Define now $Q = I - l_Q$ where l_Q here is the extension to $L^2(\pi)$. Then Q is an operator on $L^2(\pi)$ and because of the "reconstruction" process $Q1 = 1$. Extend now the codomain of l_Q to $L^2(\pi)$ by simply using the fact that $L_0^2(\pi) \subset L^2(\pi)$. Since $l_Q f \in L_0^2(\pi)$, $\forall f \in L^2(\pi)$ it follows that $(1, l_Q f) = (l_{Q^*} 1, f) = 0$. This implies $l_{Q^*} 1 = 0$ and $Q^*1 = 1$ as required.

Unfortunately nothing guarantees that condition (3) holds in general, so Q is not necessarily a Markov transition kernel. For finite state spaces condition (3) is equivalent to requiring all the entries of the transition matrix to be non-negative. It should be possible to find conditions on the spectrum of P which guarantee that requirement 3 is satisfied so that Q is indeed a proper transition kernel.

9 Comparing the performance of kernels taking CPU time into account

From a practitioner's point of view, it can be misleading to compare Markov chains in terms of asymptotic variance on a sweep-by-seep basis. This is due to the fact that different Markov chains take different amounts of time to complete one iteration, that is, to move from X_t to X_{t+1}. Furthermore, the time needed to write the computer code to implement different Markov chains can be different.

Hammersley and Handscomb [14] proposed that "the efficiency of a Monte Carlo process may be taken as inversely proportional to the product of the sampling variance and the amount of labor expended in obtaining this estimates", where the word labor is used with a very broad meaning.

Let τ_P and τ_Q be the CPU time needed to complete one iteration for transition kernel P and Q respectively. Then P is at least as efficient as Q in terms of asymptotic variance, given a fixed amount of CPU time, if

$$\tau_P v(f,P) \leq \tau_Q v(f,Q), \qquad \forall f \in L_0^2(\pi).$$

Because of (13), this condition is equivalent to

$$\tau_P(f,[2l_P^{-1}-I]f) \leq \tau_Q(f,[2l_Q^{-1}-I]f), \qquad \forall f \in L_0^2(\pi). \tag{38}$$

The comparison in (38) requires the computation of the inverse Laplacian and this is often not an easy task. Due to the multiplicative factors τ_P and τ_Q even for self-adjoint operators we cannot find a condition, equivalent to (38), that only involves the Laplacian.

The second problem that arises is related to the definition of CPU time needed to complete one iteration. In theory finding τ_P requires that an "optimal" computer program is written in order to run a Markov chain having P as its transition kernel. Since most experimental encodings are less than ideal, τ_P is very hard to measure.

It is true that from a practical point of view, the researcher is really only interested in the time per iteration of the software that s/he has available to run the Markov chains under comparison. These provide good surrogates of τ_P and τ_Q but require that the practitioner writes the computer programs for both Markov chains to be compared, unless externally provided software is available. On the other hand it is desirable that the comparison between Markov chains could be made without having to go through the extra amount of work of implementing all of them. After the comparison is made and a kernel is chosen, the computer program to run only that particular one will be written.

10 Conclusions

Given a target distribution π, there are various transition kernels that give rise to different Markov chains having π as their stationary distribution. The practitioner is often faced with the problem of choosing one of them or an efficient combination of them. Among the possible criterion that can drive this choice we focus on the asymptotic variance of the resulting MCMC estimates.

We discuss partial orderings of Markov chains with respect to this criterion; in particular we study the implications of the *Peskun ordering* [25] and propose a generalization of it the *covariance ordering*. Some of the results are extended to non-reversible Markov chains.

All the orderings we have studied are aimed to find the Markov chain that produces estimates with smallest variance in the CLT for *every* square integrable function with respect to the stationary distribution. The fact that we do not focus on a specific function is at the same time a strength and a weaknesses. If we are actually only interested in estimating the expectation of a particular function we expect that finding the Markov chain which is optimal only relatively to that specific function is desired. "Unfortunately" all our results take advantage of the condition "$\forall f \in L_0^2(\pi)$".

11 Acknowledgments

This paper is part of the first author's Ph.D. thesis which was written under the careful supervision and illuminating guide of Prof. L. Tierney. The first author's research has been partially supported trough the U. of Minnesota dissertation fellowship.

References

[1] J. S. Aujla and H.L. Vasudeva. Convex and monotone operator functions. *Annales Polonici mathematici*, 62:1–11, 1995.

[2] R. Bellman. *Introduction to Matrix Analysis.* McGraw-Hill, N.Y., 1972.

[3] J. Bendat and S. Sherman. Monotone and convex operator functions. *Transactions of the American Mathematical Society*, 79:58–71, 1955.

[4] J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, 55:25–37, 1993.

[5] K. S. Chan and C. J. Geyer. Discussion of the paper by Tierney. *Annals of Statistics*, 22, 1994.

[6] J. B. Conway. *A Course in Functional Analysis.* Springer-Verlag, 1985.

[7] P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a non-reversible Markov chain sampler. Technical Report BU-1385-M, Cornell University, 1997.

[8] N. Dunford and J. J. Schwartz. *Linear Operators Part II. First edition.* John Wiley and Sons, New York, 1963.

[9] A. Frigessi, C. Hwang, and L. Younes. Optimal spectral structure of reversible stochastic matrices, Monte carlo methods and the simulation of Markov random fields. *Annals of Applied Probability*, 2:610–628, 1992.

[10] B. Fristed and L. Gray. *A Modern Approach to Probability Theory.* Birkhäuser, Boston, 1997.

[11] A. Gelman. Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice.* Chapman and Hall, 1996.

[12] M. I. Gordin and B. A. Lifšic. The central limit theorem for stationary Markov processes. *Soviet Mathematics. Doklady*, 19:392–394, 1978.

[13] J. Größ. Some remarks on the Löwner partial ordering of Hermitian matrices. *Algebra and Stochastic Methods*, 16:191–195, 1996.

[14] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods.* Methuen, London, 1964.

[15] C. Hwang, S. Hwang-Ma, and S. Sheu. Accelerating Gaussian diffusions. *Annals of Applied Probability*, 3:897–913, 1993.

[16] J. G. Kemeny and J. L. Snell. *Finite Markov chains.* Princeton: Van Nostrand, 1969.

[17] C. Kipnis and S. R. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104:1–19, 1986.

[18] A. Korányi. On a theorem by Löwner and its connections with resolvents of selfadjoint trasformations. *Acta Scientiarum Mathematicarum*, 7:63–70, 1956.

[19] J. S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6:113–119, 1996.

[20] S. J. Liu, W. H. Wong, and A. Kong. Correlation structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society, Series B*, 57:157–169, 1995.

[21] K. Löwner. Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38:177–216, 1934.

[22] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, 1993.

[23] A. Mira and Tierney L. On the use of auxiliary variables. Technical Report 63(7–97), Univesitá di Pavia, 1997.

[24] R. M. Neal. Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. Technical Report 9508, Dept. of Statistics, University of Toronto, 1995.

[25] P. H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[26] C. R. Rao and S. K. Mitra. *Generalized Inverses of Matrices and its Applications.* John Wiley and Sons, New York, 1971.

[27] G. O. Roberts and J. S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

[28] W. Rudin. *Functional Analysis.* New York: McGraw-Hill. (Second ed.), 1991.

[29] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[30] L. Tierney. A Note on Metropolis-Hastings kernels for general state spaces. Technical Report 606, U. of Minnesota, 1995.